学者画像研究综述*

■ 王世奇[1] 刘智锋[1] 王继民[1,2]
[1] 北京大学信息管理系 北京 100871 [2]北京大学大数据分析与应用技术国家工程实验室 北京 100871

摘 要:[目的/意义]对学者画像研究进行梳理,为其相关研究提供参考。[方法/过程]通过文献调研与分析,对学者画像及其相关概念进行辨析,归纳总结学者画像的构建流程、关键技术以及主要的应用,并分析目前研究面临的挑战。[结果/结论]学者画像的构建流程包含数据搜集、数据预处理、学者标签构造与可视化分析,主要实践应用包括专家推荐、学术资源推荐和科研能力评价。当前相关研究面临多源数据获取与融合难度大、学者画像动态更新研究困难以及有效评价机制缺乏等挑战。

关键词:用户画像 学者画像 学术数据 专家推荐
分类号:G250
DOI:10.13266/j.issn.0252-3116.2022.20.008

在大数据时代,用户画像有丰富的数据资源和广泛的应用前景,逐渐成为图书情报领域的研究热点。用户画像旨在以用户为中心,从多源数据中抽取用户不同维度的信息,生成一系列标签,以刻画用户的特征。作为一种描绘用户特征、挖掘用户需求的数据服务工具,用户画像已经在图书馆、电子商务、社交媒体和智慧医疗等领域得到广泛的应用[1-4]。而伴随学术大数据的飞速增长和不断积累,如何管理和分析多源异构的学术大数据成为新的挑战。学者画像是一种以科研人员为基本单位来组织和管理学术大数据的方法。不同于一般的用户画像,学者画像的对象为科研人员,旨在从海量的学术数据中勾画出学者全貌,其在学术行为分析、学者推荐以及学者评价等研究领域具有广阔的应用前景。

学者画像的重点是刻画不同学者的特征属性,其与一般的用户画像在研究对象、数据来源、技术方法以及画像应用等诸多方面都有一定的差异。当前学者画像的相关研究,主要集中于学者画像的概念[5]、构建流程以及基于学者画像的应用研究。关于学者画像的构建方法流程研究,包含数据来源、学者画像维度设计[6]、多源数据融合技术[7-9]、学者标签的自动抽取[8]以及学者画像的可视化展示等内容;基于学者画像的应用研究,则有科研合作者推荐[6]、学者研究兴趣发现[10]等。可见,学者画像的研究已取得一定的进展。

目前国内部分学者从宏观与微观等视角对用户画像的概念、方法与技术、模型及其应用等方面的相关研究进行了归纳梳理[11-14]。但关于学者画像研究的综述,仅有袁莎等[15]在2018年从计算机领域对学者画像相关的技术、存在的问题以及未来的发展方向进行讨论和展望,缺少对于学者画像的模型构建和主要应用的总结。近年来,随着学者画像研究的发展,相关研究从数据来源、模型构建、技术方法及其应用领域等角度对学者画像领域进行了进一步探索,提出了一些新的学者画像模型和应用方向。与此同时,部分研究对学者画像的概念起源存有一些争议。因此,本研究首先对学者画像相关概念、构建流程进行梳理,并总结学者画像的主要应用方向以及面临的挑战,同时对学者画像的未来研究趋势进行展望,以期为学者画像相关研究提供参考。

为了梳理国内外关于学者画像的研究进展,本文以“学者画像”“专家画像”“科研人员画像”为检索词在知网、万方和维普学术数据库中对中文文献进行主题检索,文献分类目录限定为图书情报与数字图书馆领域、计算机软件及计算机应用,检索时间为2022年4月30日;以“scholar profile”“researcher profile”“scientist profile”为检索词在Web of Science核心库中对英文

* 本文系国家社会科学基金重点项目“开放科学数据集统一发现的关键问题与平台构建研究”(项目编号:20ATQ007)和北京大学重庆大数据研究院北京基地项目研究成果之一。
作者简介:王世奇,博士研究生;刘智锋,博士研究生;王继民,教授,博士生导师,通信作者,E-mail:wjm@pku.edu.cn。
收稿日期:2022-07-29 修回日期:2022-08-24 本文起止页码:73-81 本文责任编辑:易飞

文献进行主题检索，研究方向限定为“Information Science Library Science or Computer Science”，检索时间为 2022 年 4 月 20 日。排除不相关与重复的文献，并对检出文献的参考文献进行回溯检索，最后获取代表性的 54 篇文献作为本文综述的文献集。基于获取的文献，下文对学者画像领域的研究内容进行梳理分析。

1 学者画像与用户画像的概念辨析

目前部分研究认为用户画像的概念最早是由交互设计之父 A. Cooper 在其著作 *The inmates are running the asylum*（1999 年）中首次提出的[7,12]。但是 A. Cooper 在书中并未直接提及用户画像，与其相关的原文叙述为“A persona is a fictitious, specific and concrete representation of target users”[16]。文中提及的“persona”是指用一个虚构又独特具体的用户来代表目标用户，因此“它”通常被译作“典型用户”或者“用户角色”。Persona 与很多研究中使用数据构建的精准用户画像并不是一个概念，因此将 A. Cooper 的著作称为用户画像的起源并不准确。

另一个与用户画像起源相关的概念是“user profile”（用户简要）。A. Cooper 在其著作 *About face 3：the essentials of interaction design*（2007 年）中明确界定 persona 和 user profile 是两个完全不同的设计工具[17]。K. Baxter 等将 user profile 定义为用户属性的详细说明，如职称、经验、教育程度、关键任务、年龄范围等，是关于用户非单一的、详尽的特征集[18]。User profile 概念常被用于计算机领域的信息过滤、标签推荐系统[19]、个性化文档检索[20]等主题中。总体而言，无论是从定义上还是应用场景上，User Profile 更符合国内多数用户画像研究中对于“用户画像”一词的界定：基于海量数据，抽取用户信息并构造出用户标签集合[21]。

学者画像属于用户画像研究的一个重要分支。学者画像的研究对象为科研人员，早在 2007 年 L. Yao 等就首次提到“其区别于传统的以手动输入方式建立用户档案，他们关注如何通过整合不同类型信息之间的依赖关系，使用统一的方法自动为研究人员构建画像”[22]。之后，范晓玉等将科研人员画像定义为基于科研人员的社会属性、科研习惯与行为等信息，构建标签化的用户模型[23]。袁莎等认为学者画像是通过计算机技术自动从开放互联网中获取构建科研工作者用户模型的各维度信息，从而开展数据挖掘和应用分析过程，为专家遴选、专家推荐等具体应用提供支持[15]。与普通的用户画像不同，秦成磊等强调学者画像核心是基本信息提取、研究兴趣发现和学术影响力评估[24]。综合上述内容可以看出，学者画像在研究对象、数据来源、构建模型以及应用等方面与一般的用户画像存在一定差别。学者画像主要针对科研工作者，依托互联网上各种来源的开放数据，使用预定的规则或特定的机器学习模型来提取不同属性的信息，通过整合不同属性信息之间的依赖关系，构建精准的学者画像模型，为专家评价、专家遴选、专家推荐等应用提供支持。由于学术数据具有数据权威性高、数据规模大、数据易获取等特点，相比于一般的用户画像而言，学者画像更加精准，分析也更深入。

2 学者画像的构建流程及涉及的关键技术

学者画像构建方法与技术是该领域的研究重点。当前，有不少研究从不同视角和不同层次阐述了用户画像的构建流程。此类研究[25-26]常将用户画像的构建流程总结为搜集用户特征数据、提取用户特征信息和构建用户画像模型等步骤。学者画像在数据来源、标签体系设计以及应用等方面与普通的用户画像存在一定差异。王锐杰将学者画像构建流程总结为数据信息层、数据处理层和画像分析层，其中画像分析层中的信息分为基础维度和进阶维度两类[27]。池雪花的研究将画像构建分为个人信息描述、研究兴趣标签发现和学术影响力预测三部分，并总结了其涉及的信息抽取和文本分类关键技术[28]。范晓玉等在完成学者画像的模型表示和标签化提取后进行了学者画像可视化研究[23]。综合上述学者画像研究发现，基于大数据的学者画像构建研究框架主要包含 4 个部分，分别为学者多源数据搜集、数据预处理、学者标签构造和学者画像可视化展示，见图 1。

2.1 数据源及数据搜集方法

在学术大数据时代，大型学术数据库、学术搜索引擎与学术社交媒体平台等拥有海量的学者相关成果、行为等数据，为学者画像奠定了坚实的数据基础。学者画像的数据包含学者个人数据、科研成果数据、科研社交数据等不同类型，这三类数据的数据源见表 1。

学者个人数据指的是学者的人口统计学特征，主要包括姓名、学历、单位机构、职称等数据。此类数据主要来源于学者的个人主页、百度百科、Wikipedia 与 Aminer 平台[29]等在线网站与学术搜索引擎，其中 Aminer 平台是一个以科研人员、科技文献、学术活动三大



图 1　学者画像构建流程

表 1　学者画像数据源

数据类型	数据源	采集方法
学者个人数据	学者个人主页、百科网站、Aminer 平台等	网络爬虫、人工采集
科研成果数据	知网、万方、维普、WoS、DBLP 等	数据库导出、网络爬虫
科研社交数据	ResearchGate、Academia、学者网等	网络爬虫

类数据为基础的科技情报分析与挖掘平台，目前已包含 8 400 万科研人员数据。个人数据来源广泛，数据结构不一，给数据的采集带来较大挑战。当前，学者个人数据获取的主流方法主要包含两大类，一类是采用规则的方法从搜索引擎返回的结果或学者主页中抽取画像所需的姓名、机构、地址等信息[28]；另一类为采用机器学习算法如 CRF、LSTM 序列标注模型，对采集的主页内容进行实体识别[30]。然而，现有的基于机器学习的学者主页自动识别技术无法达到很高的识别率，仍有部分专家学者没有学者主页或者无法通过搜索引擎找到其学者主页[22]。

科研成果数据是构造学者画像的重要数据来源，主要包括学术论文、专著、科研项目和专利等数据。科研成果数据主要存在于各大学术信息库中，中文数据来源于国内的三大电子文献数据库——知网、万方和维普[31]，外文数据来源于 Web of science 等学术数据库[32]，这些数据库一般会提供一些常见格式的数据导出服务，当然也可以使用网络爬虫实现更全面灵活的数据抓取。目前互联网上有许多数字图书馆以及学术搜索引擎，例如：DBLP（Digital Bibliography & Library Project）、CiteSeerx、ACM、Google scholar 等也可作为科研成果数据的补充。此外还有存储国家自然科学基金项目和国家社会科学基金项目等项目信息的相关数据库，保存专利数据的中国专利数据库等。

学术科研社交数据是科研人员在使用一些学术社交网站时产生的关注、互动评论等数据。学术社交网站是指通过互联网帮助科研工作者进行学术研究相关的沟通交流的网站平台，与一般的社交网站不同，学术

社交网站以学术交流和学术合作为目的。目前国外比较著名的学术社交平台有 ResearchGate[33]、Academia[34]等,国内有学者网[35]、小木虫等。此外,学术成果中的合著关系也可作为学者社交数据的有效补充。

2.2 数据预处理及涉及的关键技术

学者画像的数据多源异构,需要对这些原始数据进行清洗,并对数据进行有效的融合,使之转化为适用于构建专家学者画像模型的数据。学者画像数据预处理一般包含以下三个步骤:数据清洗、信息抽取和数据融合。近年来,学者画像研究中对数据预处理有一些新的探索,本文从实践应用角度出发,总结了预处理步骤中的方法技术以及一些最新的工具。

首先,需要清洗原始数据,解决数据缺失、数据冗余的问题。对于数据缺失可使用搜索引擎检索或增加其他数据源等途径来补充,如 L. Yao 等针对没有学者主页的研究人员,从 Wikipedia 中提取其个人信息[22];M. Bravo 等选择从 DBLP 中提取出版物数据来丰富原始的大学研究人员数据[36]。对于数据冗余问题,需要删除重复数据,确保数据的唯一性。

其次,需要从文本数据中抽取所需的字段。在学者个人主页中,个人基本信息通常是非结构化的文本数据,因此需要从非结构化文本中抽取信息,而其中涉及的关键技术为命名实体识别(Named Entity Recognition,NER)。命名实体识别技术是信息抽取的重要基础,是指从自然语言文本中抽取出指定类型的实体。在学者画像领域,处理学者个人数据时主要抽取的实体包括姓名、职务、机构、邮箱等。命名实体识别方法主要有基于规则学习的方法、基于统计机器学习的方法和基于深度学习的方法。早期的命名实体识别工作大多都采用手工编写词典和规则的方法,此类型方法的优点是准确度比较高,但是查全率不高而且费时费力,语言依赖度很大,可拓展性不高。基于统计机器学习的方法将命名实体识别看作一个分类问题或者序列标注问题[37],使用人工标注分类的语料训练一些经典的机器学习分类器如 HMM、ME、CRF 和 SVM,此类方法的难点是需要大规模的训练语料以及如何构造特征工程。当前,常用深度学习的方法从学者主页等非结构化文本中进行实体识别,该方法使用词向量表示词语、字向量表示字,利用深度神经网络解决了统计机器学习方法需要构建特征工程的问题,并取得了较好的效果。目前常用的命名实体识别工具有中国科学院的 NLPIR 系统[38]、斯坦福大学的 stanza 和哈尔滨工业大学的 LTP 系统[39]等。

最后,需要整合来自不同数据源的数据,并且统一存储在数据库中进行管理。不同来源的数据对专家学者的描述标准各异,因此首先需要构建统一的元数据字段表,将不同来源的数据映射到统一的元数据字段中,实现多源数据的融合。数据融合中存在的最大挑战是专家实体消歧问题,目前已经有一些研究从语义相似度[40]、论文的特征模型相似度[41]以及结合规则和特征模型相似度[42]等角度进行了一些探索。但是从现有研究结果来看,专家学者实体消歧方法在准确率上还有一定的提升空间,未来可利用深度学习方法进一步提升实体消歧效果[43]。

2.3 学者标签构造

学者标签构造是根据学者的特征属性,对学者数据进行挖掘,抽取学者特征并用统一标准的短语对这些特征进行标识的过程,学者标签具有标准化、短文本化、语义化、专一性等特点[44]。学者画像的标签可分为个人属性标签、研究兴趣属性标签、学术能力标签以及学者社交标签四大类别。

学者个人属性标签是指描述个人基本特征的标签,例如姓名、年龄、职称和教育背景等,一般来说,个人属性标签稳定性较好,短时间内不会有较大变动。个人属性标签的来源主要是对学者个人数据的信息抽取。此外,亦可根据具体的任务,人为构造一些个人属性标签,如学术年龄、机构地理距离等[6]。

研究兴趣属性标签是学者标签体系的核心,常被用于专家推荐等场景。研究兴趣属性的标签数量、内容等没有统一的标准,常基于不同的应用场景构建不同的学者研究兴趣标签。研究兴趣标签主要来自学者的科研成果数据,最直接的方法是将学者发表论文的关键词当成兴趣标签,然而关键词数量众多且质量不一,有些关键词也并不能说明学者的兴趣主题,所以部分学者采用 LSI[45]、LDA 与 Doc2Vec[46]等算法从相关研究成果中挖掘学者的研究兴趣标签。亦有学者从网络上不同数据源中提取表示学者研究兴趣的术语,并通过 Wikipedia 整合表示研究兴趣的相关术语,以表示学者的研究兴趣[47]。此外,石湘等[48]通过梳理学者研究兴趣识别的相关文献,发现目前的学者兴趣识别研究在词汇主题层面已经比较成熟,未来的研究方向主要是网络层面的研究兴趣识别。

学者学术能力标签亦是学者画像标签体系的重要组成部分。学者的学术能力可以从学者的学术成果质量和学术影响力两个方面来反映。其中,学术成果的质量可以用论文发表期刊的级别、主持基金的级别、所

王世奇，刘智锋，王继民. 学者画像研究综述[J]. 图书情报工作,2022,66(20):73－81.

属机构权威度等指标来表示。学者的学术影响力常用 h 指数及 g 指数等衍生指数来表示[49－50]。

学者关系标签主要来自科研社交数据和科研成果数据。科研社交数据是科研人员在使用学术社交平台时产生的数据，常见的一些学术社交网站有国外的 ResearchGate、Academic.edu 和国内的科研之友、学者网等。此类网站会提供关注的学者、提问和回答的问题等社交数据。此外，可以从学者的论文等科研成果的合著关系中，获取学者的合作关系网络。

2.4 学者画像可视化分析

学者画像可视化指的是将之前构建的大量不同种类、不同权重的学者画像标签用图形等形式呈现出来，其能够直观清晰地展现和分析专家学者的各类属性。其中，构建用户信息的标签云是一种常用的可视化方法，该方法可以根据不同标签的权重，用不同大小的标签词构建词云，将用户的信息形象直观地呈现出来。现在已有很多成熟的工具可用于实现用户标签的可视化，如 WordArt、tagCloud、Tagul[51]、Tagxe-do[52] 等。此外，对于学者专家丰富的科研成果数据也可以用可视化的方法进行分析。有学者总结了几十种现有的数据可视化工具、技术以及用于分析学术数据的系统，例如使用 Pajek、Gephi 等社会网络可视化工具可以对学者的合作关系、引用关系或者关注关系进行可视化展示和社区划分等分析，或使用常见统计图表如柱状图和雷达图来展示学者的科研成果发布时间和学者的研究兴趣[53]。

3 学者画像的应用研究

3.1 专家推荐

研究兴趣是学者最重要的属性特征之一，通过学者画像中的研究兴趣等标签，并结合相似度计算等算法，可以实现相关领域专家与科研合作者推荐。在领域专家推荐方面，R. Thiagarajan 等通过使用基于本体的扩散激活网络(Spreading Activation Networks)计算用户画像之间的相似度，解决了专家发现问题[54]。胡承芳等[55]提出了基于画像技术的澜湄水资源合作领域专家库系统的设计思路，通过构建基于时空属性的人才画像模型，实现了基于澜湄合作需求的人才智能推荐功能。L. M. De Campos 等[56]通过对专家文本信息进行聚类来构建概要用户画像并提取专家感兴趣的不同隐藏主题，基于此实现了专家推荐等应用。

此外，基于学术偏好的科研合作者推荐是学者画像的另一重要应用。近年来出现了不少基于学者画像的科研合作者推荐的相关研究。这些研究主要从文献、专利以及社交媒体等渠道采集学者的学术兴趣、学术能力、合作网络[57]等信息，以构建学者画像，再通过融合学者多个维度特征的相似度进行科研合作者推荐。

3.2 学术资源推荐

学术资源信息推荐是图情领域的重要研究问题，基于学者画像进行学术资源信息推荐是当前重要的手段之一。此类推荐在图书馆领域有较多应用，通过将学者画像的兴趣、行为等特征与图书的内容匹配来进行图书资源推荐。有学者利用用户画像方法和技术，构建读者的个人画像与群体画像，并综合两者所反映出的读者借阅行为特征，实现图书的个性化推荐[58]。此后一些研究尝试融合更丰富的互联网数据，如刘海鸥等[59]提出大数据时代下基于学者画像的个性化学习资源推荐服务，结合研究人员的基本信息、研究兴趣以及社交互动数据为其提供动态的个性化资源推荐。随着学术社交媒体的不断兴起，学术社交数据日益丰富，基于学术社交数据构建学术新媒体用户画像以及基于用户画像的学术新媒体信息精准推荐模型研究亦取得一定的进展[60]。

3.3 学者科研能力评价

学者画像技术亦被应用于评估学者过去的学术能力和预测其未来的潜在能力。学者的学术能力是各大高等教育机构、研究中心和企业进行人才招聘和资助决策时主要考虑的指标，因此对其进行客观准确的评价十分重要。如韩旭等以计算机领域为例，提出了一种基于用户画像技术的学者能力指数计算及学者排名方法[61]。熊回香等将科研能力分为学术成果质量和学者的学术影响力两部分，利用一种权重主题模型表示学者的科研能力[62]。M. Lee 等[63]提出了研究人员绩效指数模型，这些模型可以衡量定性和定量绩效、研究人员影响力和增长潜力，从多个角度评估研究人员的表现，帮助他们提高研究能力。学术能力作为学者画像的重要维度之一，可以通过学者画像的学术能力等标签对学者的科研能力进行评价。

4 目前研究面临的挑战

目前已经有一批研究从概念、模型构建、关键技术以及应用等方面对学者画像进行了探索，这些研究成果为学者画像的后续相关研究提供了一定的理论和实践基础。随着开放科学运动的兴起，多源异构的科学大数据为学者画像的构建提供了丰富的数据基础，同

时也对学者画像研究带来了更大的挑战。学者画像研究面临的挑战主要包括多源数据获取与融合难度大、动态更新研究困难以及评价研究匮乏等方面。

4.1 多源数据获取与融合难度大

伴随着互联网的发展产生了海量的数据,使得构建基于多源异构数据、精准丰富的学者画像成为可能,但是如何从海量的互联网网页中筛选出所需学者网页,并实时动态地抓取和入库,是学者画像面临的挑战之一。目前常用的方法是先使用学者姓名和机构在搜索引擎中检索,再根据规则对检索结果进行筛选,然而该方法的查全率和查准率都不高[64],如何识别出所需学者个人主页仍是一个研究难点。此外,现有的研究表明存在相当一部分比例的学者没有主页或者无法通过爬虫或搜索引擎找到相关的学者主页[22],对于这部分学者的个人信息需要从其他信息源获取,例如学术出版物或者社交平台。

数据融合是学者画像的另一挑战。在多源异构数据融合过程中,存在学者重名等问题,需要使用实体消歧等技术进行解决。目前数据融合相关研究主要采用基于机器学习、基于图与基于启发式规则等方法,其中以机器学习算法为主[30]。总体而言,基于大数据的学者画像研究中的数据获取和融合技术取得了一定的进展,但是在解决大规模数据的时间复杂度和增量消歧[65]等问题上还存在很多探索空间。

4.2 学者画像动态更新研究困难

学者的相关信息处于不断变化中,为了保证学者画像的准确性和时效性,对学者不同维度的属性特征进行动态更新十分重要。然而,目前多数学者画像研究都忽略了动态更新的问题。学者画像动态更新可能的解决方式包含两类,分别是基于反馈机制的画像更新和数据驱动的画像更新。其中,基于反馈机制的学者画像更新一般采用人工的方式对学者的个人属性等信息进行修改与完善,如 AMiner 学者画像系统采用“认领 - 编辑”的方式鼓励学者本人对画像信息进行手动修改和补充[10]。这种方式可以确保更新信息的准确性,但是对于大规模的学者画像而言,更新效率太低。数据驱动的学者画像更新一般采用自动化方式更新学者的兴趣属性、学术能力属性等信息。这种基于科研成果数据的自动化更新技术需要解决实时数据搜集、构建高效的触发机制和更新机制等问题[15]。总体来看,学者画像的动态更新研究需要在学者画像的构建和应用研究基础上,投入大量的人力,存在一定的技术门槛,同时需要较长的研究周期,因此目前学者画像动态更新的相关研究较为缺乏。

4.3 学者画像评价研究匮乏

目前关于学者画像的研究大多是关注如何构建学者画像,对多源异构数据融合、信息抽取、标签组织和权重计算等流程进行创新,却少有研究从真实性、时效性和准确性等方面对已构建的学者画像模型进行科学客观的评价。目前仅有少量研究对构建的模型进行理论层面的评述[66]或针对某一领域举例说明学者画像的应用情况并进行主观评价[6],缺少基于大规模数据或者科学客观的测试样本的评价研究,这会影响学者画像模型的通用性,且不利于学者画像模型的改进。学者画像评价研究匮乏的主要原因是缺少可以用于评价的数据。学者画像评价数据包括研究人员手动标注的数据以及在学者画像应用过程中产生的使用评价、使用频率、需求匹配程度等指标数据。对于大规模的用户画像,采用人工标注的方法比较困难,只有通过深入广泛的应用后获得相关指标数据,才能用于开展学者画像系统的评价等研究。

5 总结与展望

本文对学者画像的构建与应用相关研究进行了系统梳理。首先,对学者画像的相关概念进行归纳,总结出学者画像是从海量多源异构的学术数据中,抽取出学者的不同维度的属性特征,并进行应用与分析的过程;其次,梳理了学者画像的构建流程,包含数据搜集、数据预处理、学者标签提取以及学者画像可视化分析,涉及的关键技术有学者相关实体抽取、数据融合以及可视化技术等;最后,指出学者画像相关研究在多源数据获取与融合、学者画像动态更新以及评价机制等方面面临的挑战。

针对现有相关研究中面临的挑战,有以下研究建议:关于学者画像数据获取的问题,可以使用先进的文本分类算法,例如 XGBoost、lightGBM 等算法,从搜索引擎返回的搜索结果中识别学者主页;对于学者主页中信息的抽取可使用一些大规模预训练模型进行文本表示,利用深度神经网络模型实现科研人员信息的抽取,这种不依赖人工特征的方法可以保证在各个领域有较好的通用性。此外如今有一些开放的科技大数据平台,例如:中国科学院知识服务平台[67]、粤港澳科技资

源大数据服务平台[68]等,整合了学者个人数据、文献资源数据以及社交媒体数据等不同类型的数据,可为学者画像提供数据补充。在数据融合方面,可以将现有的学术知识图谱与深度学习等方法结合,以提高学者相关数据的融合准确率。对于学者画像评价数据匮乏问题,可以通过一些关于学者画像相关的比赛,如2017开放学术精准画像大赛、CCKS 2021:AMiner学者画像大赛[69],获取专家学者画像的训练数据来进行评价研究。最后对于学者画像动态更新研究不足的问题,科研人员应该加强学者画像的实践研究,并将研究与产业紧密结合,加长研究的周期,在实践研究过程中获得学者画像的使用和反馈数据,用于支撑学者画像动态更新与评价研究。

参考文献:

[1] 许鹏程,毕强,张晗,等. 数据驱动下数字图书馆用户画像模型构建[J]. 图书情报工作,2019,63(3):30－37.
[2] 李佳慧,赵刚. 基于大数据的电子商务用户画像构建研究[J]. 电子商务,2019(1):46－49.
[3] 滕春娥,何春雨. 在线医疗社区用户画像构建与应用[J]. 图书情报工作,2021,65(12):147－154.
[4] 徐海玲,张海涛,魏明珠,等. 社交媒体用户画像的构建及资源聚合模型研究[J]. 图书情报工作,2019,63(9):109－115.
[5] 王雅娇,路佳,柯晓静. 学术画像在科技期刊中的应用研究[J]. 中国编辑,2021(4):45－49.
[6] 董文慧,熊回香,杜瑾,等. 基于学者画像的科研合作者推荐研究[J/OL]. 数据分析与知识发现 [2022－05－23]. http://kns.cnki.net/kcms/detail/10.1478.G2.20220407.1054.002.html.
[7] 范晓玉. 基于多源科技管理数据的重大项目团队成员推荐研究[D]. 西安:西安电子科技大学,2018.
[8] GENG Q, CHUAI Z, JIN J. Automatic construction of academic profile: a case of information science domain[J]. Journal of information science, 2021(4):016555152199804.
[9] HOLANDA O, ELIAS E, COSTA E, et al. Towards an agent-based approach for automatic generation of researcher profiles using multiple data sources [C]// IEEE/WIC/ACM international joint conferences on Web intelligence. New York: IEEE, 2013: 163－166.
[10] TANG J, ZHANG J, YAO L, et al. ArnetMiner: extraction and mining of academic social networks [C]// Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. Las Vegas.: ACM, 2008.
[11] 陈慧香,邵波. 国外图书馆领域用户画像的研究现状及启示[J]. 图书馆学研究,2017(20):16－20.
[12] 刘海鸥,孙晶晶,苏妍嫄,等. 国内外用户画像研究综述[J]. 情报理论与实践,2018,41(11):155－160.
[13] 宋美琦,陈烨,张瑞. 用户画像研究述评[J]. 情报科学,2019,37(4):171－177.
[14] 张海涛,徐海玲,张枭慧,等. 国内外图书情报领域用户画像研究现状及展望[J]. 图书情报工作,2019,63(7):127－134.
[15] 袁莎,唐杰,顾晓韬. 开放互联网中的学者画像技术综述[J]. 计算机研究与发展,2018,55(9):1903－1919.
[16] COOPER A. The inmates are running the asylum: why high tech products drive us crazy and how to restore the sanity[M]. 2nd ed. New York: Pearson Higher Education, 2004.
[17] COOPER A, REIMANN R, CRONIN D. About face 3: the essentials of interaction design [M]. Hoboken: John Wiley & Sons, 2007.
[18] BAXTER K, COURAGE C, CAINE K. Understanding your users: a practical guide to user research methods[M]. Burlington: Morgan Kaufmann, 2015.
[19] 许棣华,王志坚,林巧民,等. 一种基于偏好的个性化标签推荐系统[J]. 计算机应用研究,2011,28(7):2573－2575,2579.
[20] MALESZKA M, MIANOWSKA B, NGUYEN N T. A method for collaborative recommendation using knowledge integration tools and hierarchical structure of user profiles [J]. Knowledge-based systems, 2013, 47: 1－13.
[21] 张海涛,栾宇,周红磊. 用户画像:向知识迈进[J]. 图书情报知识,2020(5):131－134.
[22] YAO L, TANG J, LI J. A unified approach to researcher profiling [C]//IEEE/WIC/ACM international conference on Web intelligence. Fremont: IEEE, 2007: 359－366.
[23] 范晓玉,窦永香,赵捧未,等. 融合多源数据的科研人员画像构建方法研究[J]. 图书情报工作,2018,62(15):31－40.
[24] 秦成磊,章成志. 大数据环境下同行评议面临的问题与对策[J]. 情报理论与实践,2021,44(4):99－112.
[25] 姚远,张蕙,郝群,等. 基于本体的用户画像构建方法[C]//中国计算机用户协会网络应用分会2018年第二十二届网络新技术与应用年会论文集. 2018:232－238.
[26] 高广尚. 用户画像构建方法研究综述[J]. 数据分析与知识发现,2019,3(3):25－35.
[27] 王锐杰. 基于多源信息融合的科研学者画像及应用研究[D]. 成都:电子科技大学,2020.
[28] 池雪花. 学者精准画像的自动构建研究[D]. 南京:南京理工大学,2019.
[29] ZHANG J, TANG J. Name disambiguation in AMiner[J]. Science China (information sciences), 2021, 64(4):214－216.
[30] 牛海波,罗威,尹忠博,等. 一种基于互联网信息的开放学者画像方法:CN108090223B [P]. 2020－05－12.
[31] 昌宁,窦永香,徐薇. 基于多源数据的科技文献作者同名消歧研究[J]. 情报科学,2021,39(6):108－116.
[32] LIU G, YANG L. Popular research topics in the recent journal publications of library and information science[J]. The Journal of academic librarianship, 2019, 45(3):278－287.
[33] Research Gate[EB/OL]. [2022－05－10]. https://www.re-

searchgate. net.
[34] Academia. edu - Share research[EB/OL]. [2022 - 05 - 10]. https://www. academia. edu/.
[35] 学者网 - SCHOLAT[EB/OL]. [2022 - 05 - 10]. https://www. scholat. com/.
[36] BRAVO M, REYES-ORTIZ J A, CRUZ I. Researcher profile ontology for academic environment [C]//Science and information conference. Cham: Springer, 2019: 799 - 817.
[37] SUN J, XU J G, CEN Z W. Chinese researcher profile annotation based on conditional random fields with semantic rules [C]// World congress on engineering. v. III. : international association of engineers. London: WCE, 2011: 1818 - 1822.
[38] 张华平,商建云. NLPIR-Parser:大数据语义智能分析平台[J]. 语料库语言学,2019,6(1):87 - 104.
[39] CHE W, LI Z, LIU T. LTP: a Chinese language technology platform [C]// 23rd international conference on computational linguistics. Beijing: Coling 2010, Demonstrations,2010:13 - 16.
[40] DEMARTINI G. Finding experts using Wikipedia [C]// International conference on finding experts on the Web with semantics. Busan: CEUR-WS. org, 2007.
[41] 曾健荣,张仰森,王思远,等. 基于多特征融合的同名专家消歧方法研究[J]. 北京大学学报(自然科学版),2020,56(4):607 - 613.
[42] 朱云霞. 中文文献题录数据作者重名消解问题研究[J]. 图书情报工作,2014,58(23):143 - 148,142.
[43] 温萍梅,叶志炜,丁文健,等. 命名实体消歧研究进展综述[J]. 数据分析与知识发现,2020,4(9):15 - 25.
[44] 赵刚,姚兴仁. 基于用户画像的异常行为检测模型[J]. 信息网络安全,2017(7):18 - 24.
[45] 谢鹏. 面向学术文献的学者兴趣标签识别方法[J]. 情报工程,2019,5(3):65 - 73.
[46] 池雪花,刘丽帆,章成志. 基于学术论文的学者研究兴趣标签发现研究[J]. 情报工程,2019,5(2):28 - 39.
[47] AMINI B, IBRAHIM R, OTHMAN M S, et al. Capturing scholar's knowledge from heterogeneous resources for profiling in recommender systems[J]. Expert systems with applications, 2014, 41(17): 7945 - 7957.
[48] 石湘,刘萍. 学者研究兴趣识别综述[J/OL]. 数据分析与知识发现[2022 - 05 - 05]. http://kns. cnki. net/kcms/detail/10. 1478. G2. 20211213. 1739. 008. html.
[49] HIRSCH J E. An index to quantify an individual's scientific research output [J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102 (46): 16569 - 16572.
[50] 熊回香,杨雪萍,蒋武轩,等. 基于学术能力及合作关系网络的学者推荐研究[J]. 情报科学,2019,37(5):71 - 78.
[51] WordArt. com-Word Cloud Art Creator[EB/OL]. [2022 - 05 - 05]. https://wordart. com/.
[52] Tagxedo-Word Cloud with Styles[EB/OL]. [2022 - 05 - 05]. http://www. tagxedo. com/.
[53] LIU J, TANG T, WANG W, et al. A survey of scholarly data visualization[J]. IEEE access, 2018, 6: 19205 - 19221.
[54] THIAGARAJAN R, MANJUNATH G, STUMPTNER M. Finding experts by semantic matching of user profiles [D]. Karlruhe: CEUR-WS, 2008.
[55] 胡承芳,李季,王春芳,等. 基于画像技术的澜湄水资源合作领域专家库构建研究[J]. 长江技术经济,2021,5(6):100 - 106.
[56] DE CAMPOS L M, FERNÁNDEZ-LUNA J M, HUETE J F, et al. Automatic construction of multi-faceted user profiles using text clustering and its application to expert recommendation and filtering problems[J]. Knowledge-based systems, 2020, 190: 105337.
[57] DING Y, YAN E, FRAZHO A, et al. PageRank for ranking authors in co-citation networks[J]. Journal of the American Society for Information Science and Technology, 2009, 60(11):2229 - 2243.
[58] 何娟. 基于用户个人及群体画像相结合的图书个性化推荐应用研究[J]. 情报理论与实践,2019,42(1):129 - 133. 160.
[59] 刘海鸥,刘旭,姚苏梅,等. 基于大数据深度画像的个性化学习精准服务研究[J]. 图书馆学研究,2019(15):68 - 74.
[60] 李宇佳,王益成. 基于用户动态画像的学术新媒体信息精准推荐模型研究[J]. 情报科学,2022,40(1):88 - 93,101.
[61] 韩旭,李寒,张丽敏,等. 基于学术行为的学者排名技术及实现[J]. 电脑知识与技术,2019,15(26):1 - 3,5.
[62] 熊回香,杜瑾,代沁泉,等. 基于主题与多维计量指标的学者学术影响力评价研究[J]. 情报理论与实践,2021,44(8):22 - 27,21.
[63] LEE M, CHO M, JEONG C, et al. Researcher profiling for researcher analysis service[C]// SWCIB2014 workshop, collocated with JIST2014 conference. Chiang Mai: JIST (Workshops & Posters). 2014: 18 - 23.
[64] ZHAO J P, LIU T W, SHI J Q. Improving academic homepage identification from the Web using neural networks[C]// International conference on computational science. London: Springer, 2019: 551 - 558.
[65] 沈喆,王毅,姚毅凡,等. 面向学术文献的作者名消歧方法研究综述[J]. 数据分析与知识发现,2020,4(8):15 - 27.
[66] 胡媛,毛宁. 基于用户画像的数字图书馆知识社区用户模型构建[J]. 图书馆理论与实践,2017(4):82 - 85,97.
[67] 中国科学院知识服务平台 (las. ac. cn)[EB/OL]. [2022 - 05 - 18]. https://www. las. ac. cn/.
[68] 粤港澳科技资源大数据服务平台[EB/OL]. [2022 - 05 - 18]. https://talent. dgut-gba. cn/.
[69] CCKS 2021: AMiner 学者画像-Biendata[EB/OL]. [2022 - 05 - 18]. https://www. biendata. xyz/competition/ccks_aminer_profiling/.

作者贡献说明:

王世奇:调研与梳理文献,撰写论文初稿;
刘智锋:调研与梳理文献,修改论文;
王继民:提出研究问题和总体研究思路,修改论文。

A Review of Scholar Profiling Research

Wang Shiqi[1] Liu Zhifeng[1] Wang Jimin[1,2]
[1] Department of Information Management, Peking University, Beijing 100871
[2] National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing 100871

Abstract: [Purpose/Significance] This paper summarizes the research on scholar profile, and provides a reference for the related research. [Method/Process] Through literature research and analysis, this paper discriminated the scholar profile and its related concepts, summarized the construction process, key technologies and main applications of the scholar profile, and analyzed the challenges faced by the current research. [Result/Conclusion] The construction process of scholar profile includes data collection, data preprocessing, scholar label construction and visual analysis. The main practical applications include expert recommendation, academic resource recommendation and scientific research ability evaluation. At present, there are still some challenges in related research, such as the difficulty of multi-source data acquisition and fusion, difficulties in research on dynamic update of the scholar profile and the lack of effective evaluation mechanism.

Keywords: user profile scholar profile academic data expert recommendation